DOCUMENT RESUME

ED 170 369 TM 008 949

AUTHOR Lindvall, C. Mauritz
TITLE Basic Considerations in Assessing the Validity of Evaluation Designs.
PUB DATE [Apr 79]
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Research; *Evaluation; *Evaluation Methods; Information Utilization; *Research Design; Research Methodology; *Validity

ABSTRACT

Evaluation studies on educational questions attempt to provide answers in the form of conclusions or inferences which are derived from the information collected. Valid inferences are a result of careful research design for the study, and may be causal, descriptive, value-oriented, or probabilistic. Basic steps in designing an evaluation study are suggested: (1) developing examples of each type of inference which will result; (2) identifying major components of each inference; (3) specifying the validity concerns of each component; and (4) planning the study so that inferences will be defensible against claims of invalidity. Types of validity applicable to evaluation studies include content validity, construct validity, internal validity, population validity, and ecological validity. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170369

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Basic Considerations in Assessing the Validity of Evaluation Designs

C. Mauritz Lindvall

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C Mauritz Lindvall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Learning Research and Development Center

University of Pittsburgh

1979

The work reported herein was supported by the Learning Research and Development Center, supported in part by funds from the National Institute of Education (NIE), Department of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

TM008 949

BASIC CONSIDERATIONS IN ASSESSING THE VALIDITY OF EVALUATION DESIGNS

C. Mauritz Lindvall

University of Pittsburgh

Evaluation studies are conducted in order to arrive at defensible conclusions concerning important questions, questions which take many forms. It is now generally recognized that the term educational evaluation is used, quite legitimately, to refer to a broad range of activities associated in some general way with the assessment of the worth of some aspect of educational programs and activities. Some of the more widely discussed outlines of the various stages at which evaluation should be employed (Alkin, 1969; Stake, 1967; Stufflebeam, 1968) appear to agree that evaluation can serve such diverse functions as assessing current needs, determining the value of proposed program goals, estimating the likelihood that a planned program will achieve certain outcomes, assessing the effectiveness of the initial implementation of a new program, and describing the extent to which a program does indeed produce its desired outcomes. Of course, any student of evaluation could quickly generate an even longer list of additional functions that evaluation can serve. In serving any of these purposes an evaluation study involves an effort to arrive at defensible answers to key questions generated by these diverse functions. The needed answers take the form of conclusions or inferences from information collected as a part of the central activity of

the evaluation. A major focus of any evaluation study can then be thought of as the process of using information to make valid inferences that provide answers to the important questions. Doing the necessary planning so that the inferences from a study can be defended as having some necessary degree of validity is generally referred to as the process of designing the study.

Most evaluators come to this task of the design of the study with some knowledge of classical experimental design. They are aware of certain principles of design, have some understanding of the sources of invalidity that must be of concern in applying such designs, and probably have some knowledge of certain good and poor experimental designs. However, as they approach their task of designing an evaluation study, such persons quickly come to see that experimental designs, while very useful for studies dealing with a limited number of evaluation questions, are of little help in studies that attempt to answer many of the real questions that are posed to the evaluator. The obvious reason for this is the fact that experimental designs are intended for use in studies where the purpose is to study causal relationships and to make causal inferences. When the evaluator is concerned with making such inferences, some type of experimental design may well provide the answer to the design problem. For example, this could well be the case when one is seeking evidence concerning the hypothesis that "this program produces these results." But, as discussed above, the concern of an evaluation study is frequently something other than making this type of causal inference.

Table 1 represents an effort to suggest some of the variety in types of questions that may be of concern to an evaluator. The ten stages in planning and development and in evaluation listed in the figure are the result of the writers' effort to analyze many of the existing outlines, or "models," of

the development/evaluation process and, on this basis, to synthesize a fairly detailed sequence of steps representing a systematic development process. The figure is not an effort to develop another evaluation "model." Its only purpose is to provide a framework for a consideration of some of the variety of questions that an evaluation design must serve to answer. As an aid to this purpose the second column of Table 1 provides an example of the general question that must be of concern in an evaluation study that includes some attention to that given stage. Of course, at any given stage, it is possible that other types of questions might also be raised. As one example, the identification of promising revisions in a program (stage 8) might involve questions concerning the relative effectiveness of two alternative program components. Again, however, the purpose of Table 1 is only to suggest that evaluation does indeed involve attention to a variety of types of questions and their related inferences. An examination of the questions in the second column will show that while the inferences required at stage 10 are obviously causal inferences, the inferences required at other stages may be thought of as descriptive inferences, inferences of value, probability inferences, and others less easy to classify.

What, then, are the general design requirements if evaluators are to plan their studies so that they provide the basis for defensible conclusions concerning the exact questions that a study seeks to answer? How does an evaluator identify an appropriate design? An attempt to provide some guidance in answering this question might well start with a re-examination of the basic considerations with which Campbell and Stanley (1963) are concerned in their assessment of various designs that are appropriate for experimental

research and to point to some of the implications of parallel considerations for the design of evaluation studies.

It is proposed in this paper that just as the Campbell and Stanley explication of the key considerations in experimental design achieves its focus and direction by centering attention on what is needed if one is to be able to make valid causal inferences, so any clarification of the key considerations in developing evaluation designs will be facilitated by centering attention on what is needed if one is to be able to defend the validity of any one of the varied types of inferences needed in evaluation studies. The paper does not pretend to provide any full or detailed explication of what is involved in this approach. The hope of the writers is, however, that the basic ideas presented here have enough promise that they will be worthy of further exploration.

Validity

The concept of validity as applied to empirical inquiry in education is discussed most exhaustively in two specific contexts: (1) that of the validity of test scores or of other means of collecting and summarizing data (see Cronbach, 1971) and (2) that of the validity of the inferences that can be made from experimental research (see Campbell and Stanley, 1963). For purposes of this paper, it is important to point out that in both of these contexts the basic concern is the same, the concern for making valid inferences from given data. For example, Cronbach states that:

> "Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score. ...More broadly, validation examines the soundness of all the interpretations of a test descriptive and explanatory interpretations as well as situation-bound predictions." (Cronbach, 1971, p. 443)

Also, Campbell and Stanley open their chapter in the Handbook of Research on Teaching, by stating, "In this chapter we shall examine the validity of 16 experimental designs against 12 common threats to valid inference." (Campbell and Stanley, 1963, p. 171). The Campbell and Stanley presentation, particularly in its discussion of internal validity, focuses on one type of inference, namely causal inference. The final sentence of their chapter states that "Throughout, attention has been called to the possibility of creatively utilizing the idiosyncratic features of any specific research situation in designing unique tests of causal hypotheses" (Campbell and Stanley, 1963, p. 241). Of course, their discussion of basic considerations concerning external validity has meaning for any type of inference in which one is attempting to generalize beyond the given sample of people and situations from which data were collected. Bracht and Glass (1968) and Snow (1974) have extended this discussion of factors affecting external validity and have concerned themselves with its two basic aspects (1) population validity and (2) ecological validity. While the concerns for "test validity" may most easily be seen in the context of simple descriptive inferences (e.g., "This year's class had a higher mean score in reading comprehension than did last year's class."), they are probably relevant with all types of inference in that they speak to the question of the true meaning of the variables involved, both independent variables and dependent variables. That the general process of validation has a broad meaning that makes it relevant to the clarification and defense of all types of inferences from data is indicated by Cronbach's statement that:

> Physical scientists, one notes, speak not of validating a measuring procedure but of validation of scientific theories... . ...validation is a process for developing sounder interpretations of observations. (Cronbach, 1971, p. 443)

Planning Valid Evaluation Studies

If, as explained in the foregoing analysis, the purpose in planning and conducting an evaluation study is to enable one to make valid inferences, then the process of designing the study should begin with an examination of the types of inferences needed and then should give systematic attention to what will be necessary to enhance the validity of each such inference. This could be outlined in terms of the following four steps.

1. Develop an example of each of the types of inference that should result from the evaluation
2. Identify the major components of each inference
3. Specify the validity concerns associated with each component
4. Plan the study so that it will be possible to defend each inference, to some acceptable degree, against any claims of invalidity in an of it's major components.

The remaining sections of this paper discuss what is involved in carrying our these steps.

Develop an Example of Each of the Types of Inference That Should Result from the Evaluation

The left hand column of Table 2 provides examples of five different types of inferences that might be needed, depending upon the questions that the study is expected to answer. Statements of this type, but in a more specific form, should be formulated for each aspect of the evaluation study. This is the key step in planning and designing an evaluation so that it focuses on the important questions that must be answered. The evaluation staff must come up with statements that represent specific answers to the question "What are the types of statements that we should be able to make on the basis of our evaluation data if our study is to speak to the problems that it is intended to address?" The statements generated in this way will represent the inferences the validity of which must be supported through the evaluation design. To serve their

purposes, each of the inferences should be as specific as possible, stated in terms of the exact variables that are of concern, clearly describing the relationship or status involved, and identifying the population of persons and the settings to which it applies.

Identify the Major Components of Each Inference

The headings for columns 2, 3, 4, and 5 in Table 1 indicate the possible components of any given inference. Causal inferences will include all of these components. They state that some treatment causes a given result with certain types of subjects in a defined setting. As suggested by the examples in the table, descriptive inferences and inferences of value may well involve only one type of variable. However, all inferences describe some type of relationship or some status, that is, they specify a relationship between an independent and a dependent variable or they provide a description of status (e.g., a measure, a judgement, a verbal description) concerning one or more of the variables. Also, most inferences should refer to some type of setting and some groups of subjects, however broadly or narrowly delimited these might be.

Specify the Validity Concerns Associated With Each Component

As indicated in Table 2 the types of validity that must be of concern with respect to the independent and dependent variables can probably be classified as content validity and construct valdity as these have been defined by measurement theorists (see Cronbach, 1971). The validity of the relationship or the status described in the inference is obviously an instance of internal validity in the case of causal inferences. In the case of the other types of inference as shown in Table 2 the type of validity involved in defending the relationship or description is less easy to classify. Questions about the

validity of any generalizations concerning subjects and setting would appear to be concerns for those specific aspects of external validity that Bracht and Glass (1968) have discussed under the headings of population validity and ecological validity, repectively.

The Application of Various Types of Validity in the Design of Evaluation Studies

Since attention to all of the above categories of validity is essential to an understanding of the design considerations being proposed in this paper, it may be useful to examine each of these types of validity in terms of their application to evaluation studies.

(a) Content validity. A "slightly altered" version of the definition of content validity found in the 1966 APA edition of Standards for Educational and Psychological Tests is given in Cronbach's validity chapter:

> Content valdity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn (Cronbach, 1971, p. 444).

For the dependent variables in an evaluation study this definition would appear to be directly applicable. One must be able to show that the procedures used to assess the outcomes produced by a program do indeed provide a measure (or description) of a defensible sample of the specified universe of such outcomes.

The concept of content validity would also appear to be relevant in defending the description of an independent variable; a treatment or a program. This need for content validity would be particularly obvious in cases where some type of observation system is used to describe the extent to which the planned characteristics of a new program are indeed evident when the program is implemented. Here one should be able to defend the claim that such an observation schedule "samples the class of situations" described in the program

plan. The observations should include all relevant characteristics that are inherent in the definition of the program. Each implementation of the program should not be simply a "random sample" of the characteristics described in the program plan. Rather, each implementation should be a representative sample of the program plan (cf. Lennon, 1956; Cronbach, 1971). The program plan serves as an operational definition of the universe from which particular implementations are "sampled." Although it may be more difficult to achieve, one should also be able to defend the content validity of less objective descriptions of the characteristics of an operating program.

(b) Construct valdity. Using the same source, and with the same qualifications cited above, the following definition of construct valdity is given by Cronbach:

> Construct valdity is evaluated by investigating what psychological qualities a test measures; i.e., by determining the degree to which certain explanatory concepts or constructs account for performance on the test (Cronbach, 1971, p. 444).

Again, concern for construct validity in a dependent variable would appear to involve the same concerns that are important in describing the construct validity of any test. One must be able to defend the conceptual interpretation tendered as an explanation for the results obtained by applying a particular measurement procedure. There would seem to be a comparable concern in defending the validity of a description of an independent variable.

Any well conceived program or treatment probably represents an exemplification of certain treatment constructs (e.g., "independent study," "pupil self-direction," "positive reinforcement," "provision for pupil questions") and any procedures used to describe the program with respect to the operation of such constructs when the program is observed should be defended in terms

of the degree to which they reflect such constructs. Construct validation uses information from a variety of sources including that information obtained by content validation procedures. A program name or designation, for example, can be thought of as a short-hand designation for a cluster of constructs and when it is applied to an implementation at a particular site, one must be prepared to defend that particular implementation as at least containing a representative sample of the program elements making up the original program designation. One way of doing this is to describe all of the salient characteristics of the major components of the program and then to check to see the extent to which these are present in a particular implementation. While this is a type of content validity, the information provided is grist for construct validation as well. The elements of a program presumably are there for some pedagogical reason, frequently, for reasons stemming from educational, social psychological, or instructional theories. The conception of the original program, then, represents a nomological network Such a network can be described as consisting of "... statements connecting constructs with each other, and observable indicators with constructs..." (Cronbach, 1971, p. 476). The value of a program is explained in terms of the constructs and propositions within this nomological network.

When an evaluation is conducted it examines the validity of such a network. Cronbach's (1971) observations about the construct validation of tests appear to apply to program evaluation, as well, because when conducting evaluations contradictions occur from what was expected by the original conception of the program.

> When a contradiction of expectation occurs, one challenges parts of the network by suitable further studies until he "locates the trouble." The resolution may take many forms (some of which are the following)...

1. The interpretation of a supposed indicator maybe declared invalid...
2. A construct may be abandoned or two constructs may collapse into one...
3. The construct may be subdivided and new indicators developed to distinguish between new constructs...
4. The network maybe modified to account for the anomaly. Most often, the construct name and the indicators remain, but the set of sentences that outline associated characteristics is modified (Cronbach, 1971, p. 477).

The details and categories of evidence needed for the construct validation of a program are too numerous to specify in this paper. It should be noted, however, that the evidence comes from three board groups of procedures: logical analysis, correlational studies, and experimental studies (Cronbach, 1971). The evaluator is referred to the Cronbach chapter for ideas that might be adapted to evaluation research.

(c) Internal validity. In their extended and classic discussion of the concept of internal validity, Campbell and Stanley (1963) describe the many factors with which a researcher must be concerned if one is to be prepared to defend the validty of any causal inference against other seemingly plausible explanations for results obtained. If any given evaluation study is carried out to investigate causal claims, the same concerns presented by Campbell and Stanley must be dealt with in designing that evaluation. If the evaluator wishes to be able to state that "this program has been shown to produce these outcomes," then he or she must be in a position to defend this causal claim against other competing explanations of how the data are to be interpreted. It should be noted that this line of inquiry is also analogous to construct validation, particularly to what has been refered to as convergence of indicators and discriminability from other constructs (Campbell and Fiske, 1959; Cronbach, 1971).

(d) Types of Validity Associated with "Status" Description. Although the writing on internal valdity has been focused on questions of the validity of causal inferences arising from empirical inquiry, it is proposed here that something comparable must be of concern in those inquiries where other types of inferences are made. To explore this we might note the first example of a descriptive inference as given in Table 2 which states that "First graders in this school are below the national average in arithmetic problem solving." With this inference one must, of course, be prepared to defend the content and construct validity of the measure of arithmetic problem solving. But the "status" part of the inference states that these students are "below the national average." Here, a critic of this inference might question the representativeness of the norm group. A study designed to make this inference must use a normed test for which the norm group can really be defended as representative of a relevant group. An evaluation occuring in a particular school district, for instance, might find it more relevant to use "large city schools" or "New Jersey schools" as the appropriate norm group, rather than a "national" group--even though it can be demonstrated that the latter norm group is on the whole quite representative of the nation's schools. The evaluator's problem is that when such specialized norm groups are defined, norm data about them are seldom available.

The second example of a type of descriptive inference given in Table 2 states that "Program X has characteristics Y when observed in setting B." To make this type of inference the evaluator must be as certain as possible that the observations involved are based on observations of a real exemplification of Program X. To do this the evaluation report must provide information to show that this implementation of Program X has content and construct valdity,

that it shows the qualities that serve to define Program X. But the report must also defend the valdity of the part of inference that says "Has characteristics Y." To make this example a little more specific, let us assume that such an inference says that "Program X is characterized by rather general pupil attentiveness to lesson activities. ." The validity of this inference will depend not only upon the validity of the measure of pupil attentiveness but also on the extent to which it can be shown that this attentiveness is rather typically true of this implementation of Program X. Although one may legitimately choose to think of this quality as an aspect of the reliability (stability) of the attentiveness measure, it is still an essential consideration in designing a study so that the desired inference may be defended as valid.

The example of a "value inference" shown in Table 2 states that "The relative value of these goals for the new curriculum as judged by our teachers is _____." It is assumed that this inference statement would be completed by listing the program goals in the order of their judged value along with a specific value weight assigned to each. As indicated in the table, one must first be concerned with the construct valdity of the goal statements. Without taking too much liberty with the accepted use of the term construct validity it is here proposed that a specific program goal can be thought of as one proposition in the broader network of propositions that taken together describe the basic purpose(s) of a program or organization. As such, this goal needs to be defended by construct validation procedures that lead to statements about whether the goal can fit into the nomological network in a way that demonstrates its relevance for achieving progress toward the basic purpose(s) of the program or organization. The question that is addressed is "Would achievement of the program goal imply that progress has been made toward the related basic purpose(s)?"

A second aspect of the value inference is the phrase "the relative value ... as judged by our teachers." In defending the validity of this part of the statement one must be able to defend the procedure used for soliciting judgements of value as well as the context in which such judgements occur. Are the personnel who operate the program judging this goal to be of high value because it is something that they know how to accomplish with existing levels of effort? Are judges seriously willing to entertain competing goals that might require considerable new effort, retraining, or change in the "standard operating procedure"? This is, again, a matter of defending the content and construct validity of this measure of value. This is essential if the evaluator is to be in a position to defend the inference concerning the value "status" of the goals.

The final example of a type of inference listed in Table 2 is a "probability inference" stating that "Program X, as outlined in this plan, has (a specified) probability of achieving its defined goals." This type of inference, which is essential when one makes an evaluation of plans rather than of implemented program, might be thought of as a causal inference in which empirical evidence concerning causal claims is replaced by subjective estimates of the probability that a program will produce certain results. As the table indicates, the person making this type of inference must be concerned with the validity of both the independent and dependent variables. The label "Program X" must refer to a program having clearly defined components and qualities and representing the application of certain theoretical principles. The concerns for the validity of the goals will be the same as those discussed above in conjunction with the example of a value inference. The validity of the "has ... probability" part of the inference will have to be defended in terms of the

validity of the method used in determining liklihood or probability. This, again, might be considered a concern for construct validity since one must defend the fact that the procedure provides a measure of the construct "probability."

Population and Ecological Validity. Campbell and Stanley (1963) discuss threats to external validity as being those aspect of the research procedure used which would limit the extent to which one could take the findings and conclusions obtained from one's sample and generalize these to some larger population of subjects and variety of of settings. Bracht and Glass (1968) have expanded on some of the important considerations associated with this issue and have divided it into concerns for population validity and for ecological validity. The first of these categories covers concerns for the types of persons to which conclusions may apply, the second to the types of settings and conditions in which the conclusions may be applicable. In most types of empirical inquiry (research, evaluation, etc.) one must work with some limited sample of subjects and settings but wishes to generalize to some larger domain. It would appear that in all evaluation situations one must give careful attention to the level of generalizing that is required. Also, it is likely that all of the threats to population and ecological validity, as discussed by Campbell and Stanley (1963), Bracht and Glass (1968), and Snow (1974), must be considered.

Plan the Study So That It Will Be Possible to Defend Each Inference, to Some Acceptable Degree, Against Any Claims of Invalidity In Any of It's Major Components.

As stated earlier, this paper offers no simple answer to the question of how to design an evaluation study. The basic reason for this is that the types

of inferences that one needs to make from an evaluation study are so varied that there is no simple set of designs from which one may chooose to get the appropriate one for any given study. The design requirements of most evaluation studies are relatively unique and must be met by a careful examination of the validity demands associated with both the independent and dependent variables, the reported relationships or status, and the populations and setting included in each inference.

In introducing the concept of a "patched-up" design, Campbell and Stanley (1963, p. 227) suggest that "one starts with an inadequate design and then adds specific features to control for one or another of the recurrent sources of invalidity." It might be suggested that this is the rather typical procedure that must be followed in designing most evaluation studies. There are few simple and elegant designs that can be adopted. Rather, one must face the task of taking a first approximation to the design needed and then going through a careful "patch-up" process that will result in maximum validity in the inferences and conclusions that the study is to produce.

Summary

This paper represents an attempt to outline the considerations that must be of basic concern as an evaluator attempts to design a given evaluation study. It points out that the guidance given by Campbell and Stanley (1963) in their classic work on experimental design is directed toward permitting the researcher to defend the validity of any causal inferences derived from the research study. It suggests similarly, that, the purpose in designing evaluation studies is to permit one to make valid inferences of the type re-

quired to answer the questions with which the evaluation is concerned. In evaluation studies the needed inferences will at times be causal inferences but will frequently be inferences of other types. This paper directed some attention to some of these types: descriptive inferences, value inferences, and probability inferences. In designing studies that will result in inferences of these various types, the evaluator must attempt to plan the study so that the validity of each inferences can be defended. This paper has proposed that to achieve this, the design of an evaluation study must give attention to content and construct validity, as these concepts have been defined by measurement specialists, and to internal validity, population validity, and ecological validity, as these concepts have been defined by persons attending to the requirements of experimental research. It is suggested that the basic steps in the design of an evaluation study are:

1. Develop an example of each of the types of inference that should result from the evaluation.
2. Identify the major components of each inference.
3. Specify the validity concerns associated with each component.
4. Plan the study so that it will be possible to defend each inference, to some acceptable degree, against any claims of invalidity in any of its major components.

It is obvious that these steps and the expansion on them provided in this paper do not result in any simple formulas for selecting the proper design for an evaluation study. They should, however, suggest the possible

sources of invalidation of inferences and conclusions which one must guard against in planning the study. To paraphrase the concluding sentence from the Campbell and Stanley (1963) chapter quoted earlier, the answer to the evaluation design problem will probably be found in "creatively utilizing the idiosyncratic features of any specific evaluation situation in designing unique conditions that permit valid inferences of the type required."

References

Alkin, Marvin C., "Evaluation Theory Development," Evaluation Comment, University of California, Los Angeles, 1969, 2(1), 2-7.

Bracht, G. H., & Glass, G. V. "The external validity of experiments." American Educational Research Journal, 1968, 5 (4) 437-474.

Campbell, D. T., and Fiske, D. W., "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological Bulletin, 1959, 56, 81-105.

Campbell, D. T., and Stanley, Julian C. "Experimental and Quasi-Experimental Designs for Research on Teaching," N.L. Gage, (Ed.). Handbook of Research on Teaching, Chicago: Rand McNally and Co., 1963

Cronbach, L. J. "Test Validation," Chapter 14 in R. L. Thorndike, ed., Educational Measurement, 2nd ed. (Washington, D.C.: American Council on Education, 1971).

Lennon, R. T., "Assumptions Underlying the Use of Content Validity," Educ and Psych. Measurement, 1956, 16, 294-304.

Snow, Richard E. "Representative and Quasi-Representative Designs for Research on Teaching." Review of Educational Research, 1974, V. 44, No. 3, 265-292.

Stake, Robert E. "The Countenance of Educational Evaluation. Teachers College Record, 68: 523-40; 1967.

Sutfflebeam, Daniel L., et al. Evaluation and Decision Making in Education. Itasca, Illinois: F. E. Peacock Publishers, Inc., 1970.

Table 1. Stages in a Comprehensive Planning and Development Process Showing Evaluation Activities Appropriate for Each Stage

Stage in Planning and Development	Evaluation Activities (Question to be Answered)
1. Identify basic purposes and values sought by this entity (school, agency).	1. Edit and clarify purposes. Determine perceived worth. (What are the basic purposes of this entity?)
2. Select those basic purposes that represent areas of current need.	2. Determine present level of achievement of each basic purpose. (What important purposes are not being achieved?)
3. Specify goals of new program or activity or goals of effort to revise existing program.	3. Assess clarity and quality of statements. Determine perceived worth. (Is development directed toward clear and worthwhile goals?)
4. Develop rationale concerning promising ways of achieving goals.	4. Examine rationale for pertinence, comprehensiveness, and its empirical basis. (Is development based on a defensible approach?)
5. Outline general plan of program for achieving goals.	5. Assess clarity, specificity of plan, its relationship to goals and to rationale. (Does plan outline a feasible program for achieving goals?)
6. Produce all materials necessary for implementation of program.	6. Assess completeness and quality of materials. (Are materials necessary for implementation completed?)
7. Implement planned program.	7. Assess validity and feasibility of implementation. (Can this program be implemented as planned?)
8. Revise program on basis of information obtained from implementation.	8. Determine effectiveness of operating program. (How can this program be improved?)
9. Package the completed program.	9. Describe the operating characteristics of completed program. (What things actually take place when this program is operating?)
10. Describe the results produced by the program.	10. Assess achievement of intended and unintended outcomes (What results does this program produce?)

Table 2. Types of Validity for Which Provisions Must be Made in Designing Evaluation Studies That Will Permit Valid Inferences, Showing Applications to Selected Examples of Types of Inferences

Examples of Types of Inferences Used in Evaluation Studies	Validity Concerns Associated with Each Type of Variable or Other Component of Given Inference				Essential Considerations in Design
	Independent Variables	Dependent Variables	Relationship or Status Involved	Subjects and Setting	
Causal Inferences. "Program X produces outcome Y with students of type A in setting B."	Content and construct validity of the defined program.	Content and construct validity in measures of outcomes	Causal relationship. Internal validity considerations as in Campbell, Stanley	Population and Ecological validity considerations as in Bracht, Glass	Valid and documented implementation. Valid outcome measures. Causal design (e.g., experimental design). Defined setting and sample.
Descriptive Inferences. "First graders in this school are below national average in arithmetic problem solving."		Content and construct validity in measure of arithmetic problem solving.	Description. Validity of test norms must be defended.	(as above)	Valid sample of 1st graders. Valid measure of problem solving. Valid norms.
"Program X has characteristics Y when observed in setting B."	Content and construct validity of the defined program.		Description. Validity of observation.	(as above)	Valid measures of characteristics Y. Defined setting and sample.
Value Inferences. "The relative value of these goals for the new curriculum as judged by our teachers is_____."		Construct validity of the statement of the goals.	Description of value. Validity of assessment of value must be defended.	(as above)	Valid sample of teachers. Valid procedure for collecting judgements.
Probability Inferences. "Program X has this probability of achieving its defined goals."	Content and construct validity of the defined program.	Construct validity of the statement of the goals.	Description of probability. Validity of method for estimating probability must be defended.	(as above)	Defensible group of judges. Appropriate procedure for assessing probability. (e.g., subjective and/or data based).